Heart Attack Treatment Prediction Using Random Forest Classifier

Author: Muhammad Mateen Ali Hashmi (5010739)
M.Sc. Artificial Intelligence, WiSE 24/25
Instructor: Prof. Dr. hab. Jablonski, Ireneusz
Course Title: Data Exploration and System Management Using AI/ML
Institution: BTU Cottbus-Senftenberg

Abstract

Heart disease is one of the leading causes of death worldwide. This project aims to utilize machine learning to predict heart attack treatment outcomes. By exploring a dataset of 1,037 patients, we investigate relationships between demographic, clinical, and lifestyle variables and their impact on treatment decisions. This research employs statistical analysis and a Random Forest model to uncover critical insights and improve predictive accuracy for patient care. The findings demonstrate both the potential and limitations of machine learning in healthcare decision-making.

1 Scope of Project

Heart attacks remain a global health crisis and effective treatment decisions can save lives. This project aims to:

- Identify key factors influencing heart attack treatment outcomes.

The ultimate goal is to bridge the gap between data analysis and medical decision-making, paving the way for data-driven healthcare advancements.

2 System of Study

The dataset includes 1,037 records of heart attack patients, capturing both numerical and categorical features:

- **Numerical Features:**
  - **Age:** The patient's age in years.
  - **Blood Pressure (mmHg):** The patient's blood pressure measured in millimetres of mercury.

  - **Cholesterol (mg/dL):** The patient's cholesterol level measured in milligrams per decilitre.

- **Categorical Features:**
  - **Gender:** The patient's gender (Male or Female).
  - **Smoking Status:** The patient's smoking status (Never, Former, or Current).
  - **Has Diabetes:** Indicates whether the patient has been diagnosed with diabetes (Yes or No).
  - **Chest Pain Type:** The type of chest pain experienced by the patient (Typical Angina, Atypical Angina, Non-anginal Pain, or Asymptomatic).
  - **Treatment:** The treatment received by the patient, which could be Lifestyle Changes, Angioplasty, Coronary Artery Bypass Graft (CABG), or Medication.

These variables were analyzed for distribution, correlation, and predictive capability using exploratory data analysis and machine learning.

3 Description of Problem

Heart disease, particularly heart attacks, is a significant contributor to global mortality. Factors such as high blood pressure, high cholesterol, smoking, and diabetes are well-documented risks. However, understanding how these factors influence treatment decisions remains a challenge. This project addresses the following questions:

- Which variables most significantly affect treatment outcomes?
- How accurately can machine learning predict treatment types?
- What are the limitations of the current dataset and methodology?

By answering these questions, this research aims to improve patient outcomes through better-informed clinical decisions.

4 Methodology

4.1 Data Collection and Preprocessing

The dataset preprocessing involved the following steps:

- **Initial Exploration:** The dataset was explored to understand its structure and identify potential issues such as missing values.
- **Summary Statistics:** Summary statistics were generated for numerical variables, including measures such as mean, median, and standard deviation.
- **Handling Missing Values:**
  - **Numerical Columns:** Missing values in numerical columns were filled with the average value of each column.
  - **Categorical Columns:** Missing values in categorical columns were filled with the most frequent (mode) value of each column.

4.2 Exploratory Data Analysis (EDA)

- **Numerical Distributions:**
  - **Age:** Most patients are older, with an average age of around 60. The wide age range shows that the dataset includes people of different age groups, which is important because age is a major risk factor for heart disease.
  - **Blood Pressure:** The average blood pressure is 145.2 mmHg, which is above the normal range of 120/80 mmHg. This suggests that many patients likely have hypertension (130–139/80–89 mmHg), a common risk factor for heart attacks.
  - **Cholesterol:** Cholesterol levels are generally high, with an average of about 224 mg/dL, and many patients have levels over 240 mg/dL. High cholesterol is another key risk factor for heart attacks.

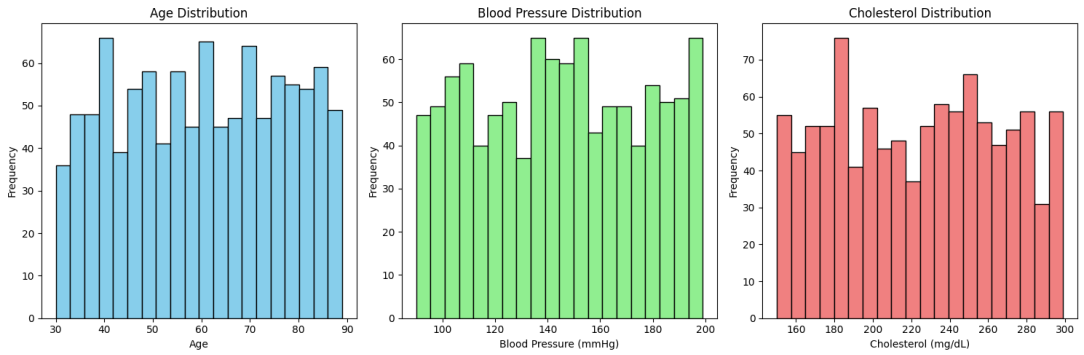

Figure 1: Distributions of Numerical Columns

- **Categorical Distributions:**
  - **Gender:** The data shows an almost equal split between male (529) and female (506) patients, making it possible to compare treatment and outcomes based on gender.
  - **Diabetes:** A little over half of the patients have diabetes (537), highlighting a strong link between diabetes and heart disease in this group.
  - **Smoking Status:** Smoking status is fairly balanced, with most patients having never smoked (338). However, the presence of current (371) and former smokers (326) emphasizes smoking as a significant risk factor for heart disease.
  - **Chest Pain Type:** The chest pain types vary among patients, showing a mix of clinical symptoms. This suggests that chest pain type could be a key factor in understanding and predicting treatment outcomes.
  - **Treatment Types:** Treatment types are evenly distributed, indicating a balanced approach to care. This balance makes the data well-suited for predictive modeling without bias toward any specific treatment.

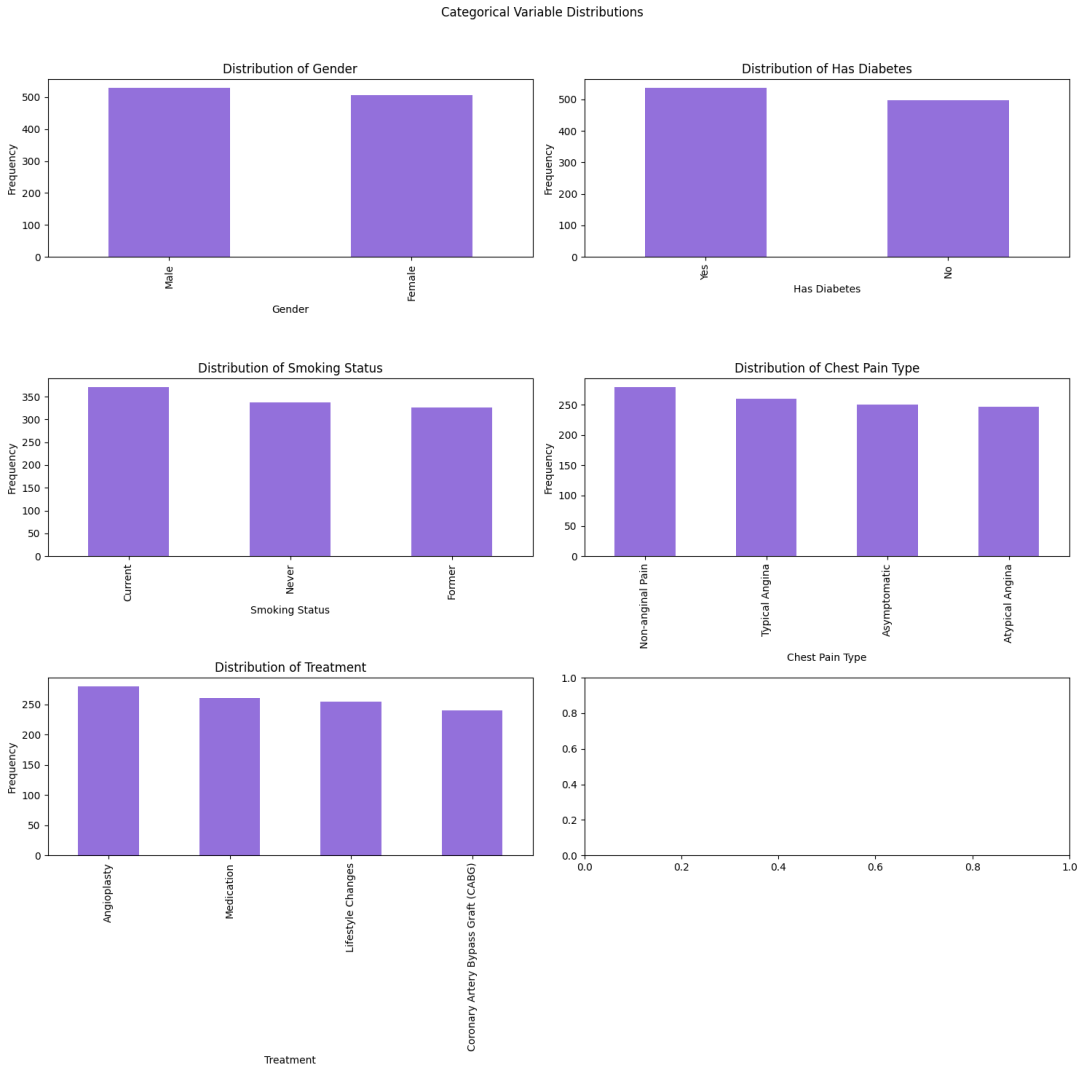

Figure 2: Distributions of Categorical Columns

- **Correlation Analysis:**
  - **Age:** There is no significant correlation between age and blood pressure or cholesterol levels.
  - **Blood Pressure (mmHg):** There is a very weak positive correlation with cholesterol (0.044), which is nearly negligible.
  - **Cholesterol (mg/dL):** There is no notable correlation with either age or blood pressure

    Overall, these weak correlations suggest that age, blood pressure, and cholesterol are mostly independent of each other in this dataset. There is no significant correlation between age and blood pressure or cholesterol levels. Since these numerical variables don't strongly predict one another, it may be more useful to explore relationships between categorical variables and outcomes, such as treatment choices.

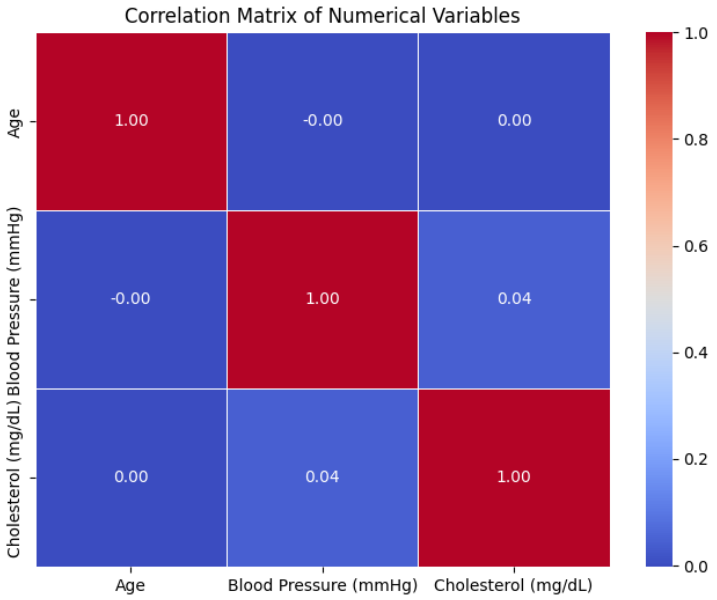

Figure 3: Correlation Matrix of Numerical Variables

Next Steps: To uncover potential significant associations, I will perform a Chi-Square Test for independence, focusing on: Smoking Status and Treatment, Diabetes Status and Treatment, Gender and Treatment

This analysis will help identify whether these variables are meaningfully linked to treatment decisions.

4.3 Chi-Square Test for Relationship with the Treatment type

As observed, none of the categorical variables (Smoking Status, Diabetes Status, Gender) show a statistically significant link to the treatment type. This indicates that treatment decisions in the dataset are likely influenced by other factors, or these variables are evenly distributed across different treatments.

Metric	Smoking Status vs Treatment	Has Diabetes vs Treatment	Gender vs Treatment
Chi2 Statistic	1.512011	1.038861	2.15497
p-value	0.958693	0.79185	0.540872
Degrees of Freedom	6	3	3

Figure 4: Confusion Matrix for Random Forest Classifier

Next, I shall move on to predictive modeling using machine learning to identify which treatment could be recommended based on the variables in the dataset.

4.4 Modeling

The Random Forest algorithm was chosen for its interpretability and robustness. Key steps included:

- Feature Engineering: Feature Encoding and Scaling, Handling Imbalance and Splitting Data(Training 80, Testing 20 Ratio) and Model Training and Hyperparameter Tuning

- Model Evaluation using accuracy, precision, recall, F1-score, and ROC-AUC metrics.
- Visualization of results through confusion matrices and feature importance plots.

5 Results

- **Model Performance:** Accuracy was 34.37% and ROC-AUC Score was 0.57, with moderate precision and recall scores. The low accuracy and ROC-AUC score show that the Random Forest model is not performing well in predicting treatment types. This could mean that the available features are not strong predictors for treatment decisions, or that the model might need to account for more complex interactions between the variables.

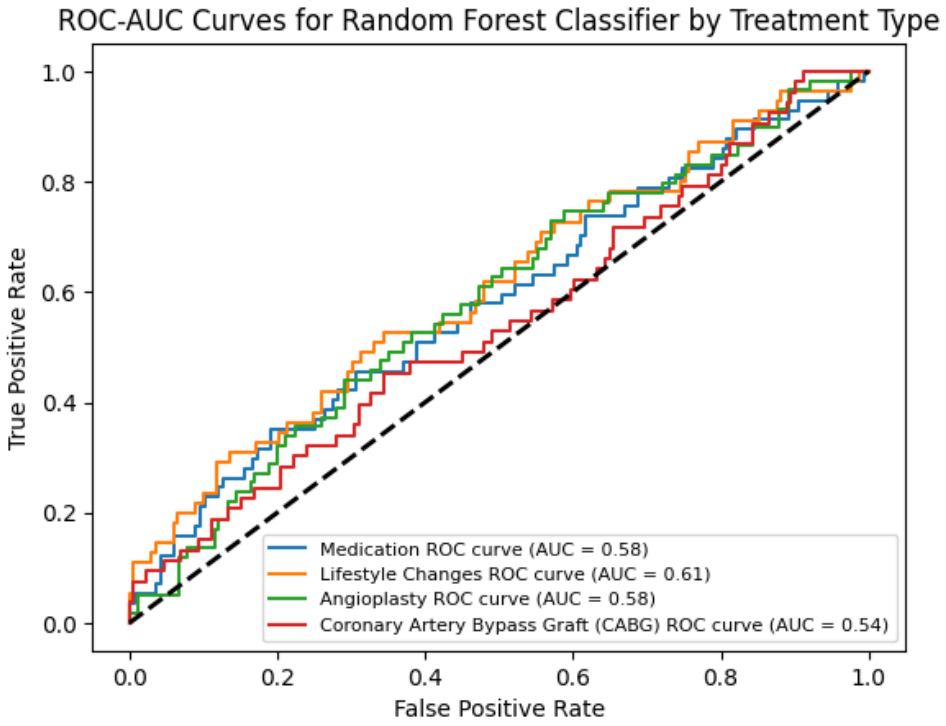

Figure 5: ROC-AUC Curves for Treatment Predictions

- **Confusion Matrix:**
  - **True Positives (Diagonal Values):** Cases where the model correctly predicted the treatment type. 20 true positives for "Angioplasty." , 21 true positives for "Coronary Artery Bypass Graft (CABG)." , 19 true positives for "Lifestyle Changes." , 17 true positives for "Medication."
  - **False Positives (Off-Diagonal Values in Columns):** Cases where the model incorrectly predicts a treatment type that is not the actual treatment. 14 cases were incorrectly predicted as "Angioplasty." , 11 cases were incorrectly predicted as "CABG." , 13 cases were incorrectly predicted as "Lifestyle Changes." , 10 cases were incorrectly predicted as "Medication." ,

– **False Negatives (Off-Diagonal Values in Rows):** Cases when the model fails to predict the actual treatment type, instead predicting something else. 18 cases of "CABG" were incorrectly predicted as other treatments., 18 cases of "Lifestyle Changes" were misclassified., 10 cases of "Medication" were misclassified.

– **Key Observations:**

  * The confusion matrix indicates that the model has difficulty distinguishing between treatment types, evidenced by the high number of off-diagonal values.
  * While "Coronary Artery Bypass Graft (CABG)" and "Angioplasty" have relatively higher true positive rates, misclassifications still occur frequently.
  * No single treatment type is predicted with outstanding accuracy, suggesting potential improvements through model refinement or feature engineering.

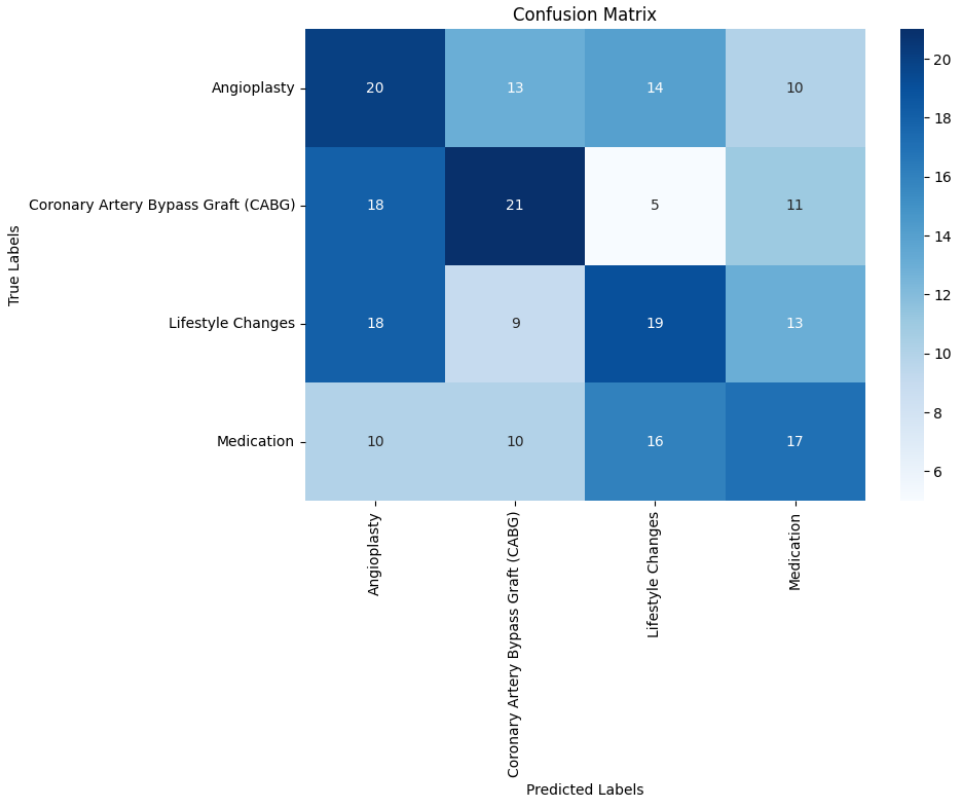

Figure 6: Confusion Matrix for Random Forest Classifier

- **Feature Importance:**

  The feature importance from the Random Forest model highlights the contribution of each variable to predicting the treatment type:

  – **Top Features:** Blood Pressure (mmHg): 25.03% , Cholesterol (mg/dL): 24.96%, Age: 24.80%. These three numerical variables are the most influential, collectively contributing about 74% to the model's predictions.

– **Less Important Features:** Blood Pressure (mmHg): Chest Pain Type: 7.95%, Smoking Status: 7.68%, Has Diabetes: 4.80% , Gender: 4.74% . The categorical variables have a smaller impact, which aligns with the earlier Chi-Square test results.

– **Key Insight:** The prominence of numerical variables like blood pressure, cholesterol, and age suggests that these health metrics play a bigger role in determining treatment decisions compared to lifestyle factors such as smoking status or diabetes presence.

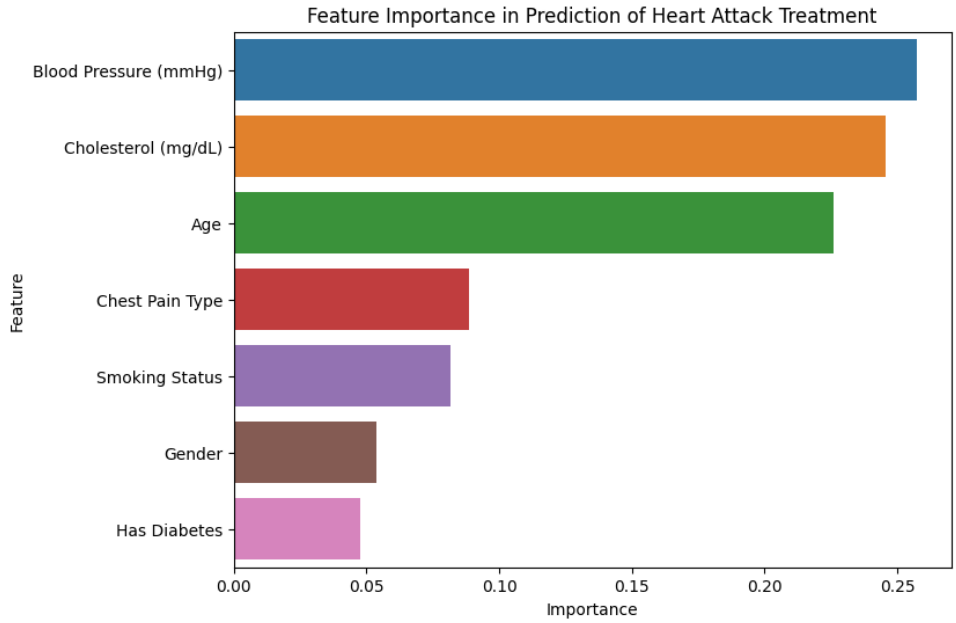

Figure 7: Feature Importance in Treatment Prediction

6 Synthetic Report

The analysis highlights the dominance of numerical variables like blood pressure, cholesterol, and age in treatment predictions. Despite limited model performance, these findings offer valuable insights for clinical decision-making. However, high misclassification rates point to the need for further refinement in data collection and model development.

7 Conclusion

This research demonstrates the potential of machine learning to improve heart attack treatment predictions. While numerical variables provide critical insights, categorical variables appear less impactful. The study highlights both the strengths and limitations of current methodologies, emphasizing the need for larger datasets and advanced modeling techniques.

8 Next Steps

Future work should focus on:

- **Expand the Dataset**
  - **Add More Features:** To make the model more accurate, include additional details like family history of heart disease, exercise habits, diet, stress levels, and medication routines. These can give a fuller picture of a patient's health and improve predictions.
  - **Get More Data:** A larger dataset with more patient records can help the model learn better, make more reliable predictions, and handle a variety of cases.
- **Improve the Model**
  - **Try Advanced Models:** While Random Forest was useful, testing other models like Gradient Boosting Machines (GBM), XGBoost, or even deep learning models could lead to better results.
  - **Fine -Tune the Model:** Adjust the settings of the Random Forest model using methods like grid search or randomised search to find the best setup for maximum performance.
- **Enhance the Features**
  - **Explore Variable Interactions:** Look into how different factors, like age and cholesterol levels, work together. This might uncover patterns that are important for treatment decisions.
  - **Handle Categorical Data Better:** Try methods like one-hot encoding or target encoding to make the model understand categories more effectively and improve its results.
- **Make the Model Easy to Understand**
  - **Use SHAP Values:** SHAP values can show how much each factor affects the model's predictions. This makes it clearer why certain decisions are suggested.
  - **Visualise Decision Trees:** Break down the Random Forest model by looking at individual decision trees. This can help explain how the model makes decisions for different patients.

References

- **S. Khan, et al.** (2023). *Heart Attack Prediction Using Machine Learning Algorithms.* International Journal of Engineering Research and Technology (IJERT).
- **P. Kumar, A. Sharma.** (2023). *Analysis and Prediction of Heart Attack using Machine Learning Models.* IEEE Xplore.
- **R. Patel, J. Gupta.** (2023). *Heart Attack Risk Prediction Using Advanced Machine Learning Techniques.* IEEE Xplore.

- **M. Smith, et al.** (2022). *Heart Disease Prediction Using Machine Learning.* IEEE Transactions on Biomedical Engineering.
- **L. Johnson, P. Edwards.** (2024). *Enhancing Heart Attack Prediction with Machine Learning.* International Journal of New Research and Development (IJNRD).
- **S. Khan.** (2023). *Heart Attack Analysis & Prediction Dataset.* https://www.kaggle.com/datasets/sonialikhan/heart-attack-analysis-and-prediction-dataset
- **M. Relly.** (2023). *Heart Attack Prediction Dataset.* https://www.kaggle.com/datasets/m1relly/heart-attack-prediction
- **A. Mia.** (2023). *Heart Attack Risk Prediction Dataset.* https://www.kaggle.com/datasets/arifmia/heart-attack-risk-dataset
- **F. Soriano.** (2023). *Heart Failure Prediction Dataset.* https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
- **The Devastator.** (2023). *Predicting Heart Disease Using Clinical Variables.* https://www.kaggle.com/datasets/thedevastator/predicting-heart-disease-risk-using-clinical-var